Greedy Geometric Optimization Algorithms for Collection of Balls

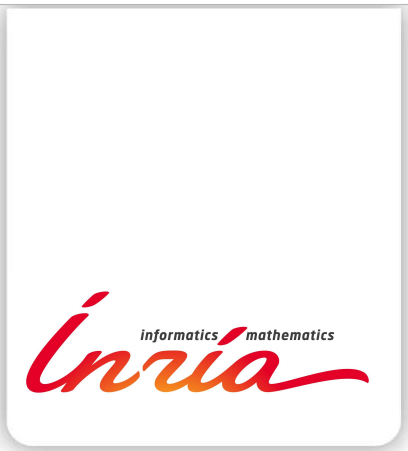
Frédéric Cazals, Tom Dreyfus, Sushant Sachdeva, Shah Nisarg

Greedy Geometric Optimization Algorithms for Collection of Balls

F Cazals and T. Dreyfus and S. Sachdeva and N. Shah

RESEARCH REPORT
N° 8205
January 2013
Project-Team ABS

ISSN 0249-6399 ISRN INRIA/RR--8205--FR+ENG



Greedy Geometric Optimization Algorithms for Collection of Balls

F Cazals * and T. Dreyfus † and S. Sachdeva ‡ and N. Shah §

Project-Team ABS

Research Report n° 8205 — January 2013 — 23 pages

Abstract: Modeling 3D objects with balls is routine for two reasons: on the one hand, the medial axis transform allows representing a solid object as a union of medial balls; on the other hand, selected shapes, and molecules in particular, are naturally represented by collections of balls. Yet, the problem of choosing which balls are best suited to approximate a given shape is a non trivial one. This paper addresses two problems in this realm.
The first one, *conformational diversity selection*, consists of choosing k molecular conformations amidst n, so as to maximize the *geometric diversity* of the k conformers. The second one, *inner approximation*, consists of approximating a molecule of n balls with $k < n$ balls. On the theoretical side, we demonstrate that for both problems, a geometric generalization of max k-cover applies, with weights depending on the cells of a surface or volumetric arrangement. Tackling these problems with greedy strategies, it is shown that the $1 - 1/e$ bound known in combinatorial optimization applies in some cases but not all. On the applied side, we present a robust and effective implementation of the greedy algorithm for the inner approximation problem, which incorporates the calculation of the exact Delaunay triangulation of a points whose coordinates are degree two algebraic number, of the medial axis of a union of balls, and of a certified estimate of the volume of a union of balls. In particular, we show that the inner approximation of complex molecules yields accurate coarse-grain models with a number of balls 100 times smaller than the number of atoms, a key requirement to simulate crowded protein environments.

Key-words: Geometric optimization, geometric approximation, collection of balls, inner approximation, molecular conformations

* Inria Sophia-Antipolis; frederic.cazals@inria.fr
† Inria Sophia-Antipolis; tom.dreyfus@inria.fr
‡ Princeton University; sachdeva@cs.princeton.edu
§ Carnegie Mellon University; nisarg89@gmail.com

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Algorithmes Gloutons pour des Familles de Boules

Résumé : Les boules jouent un rôle central en modélisation géométrique pour deux raisons: d'une part la transformée associée à l'axe médian permet de représenter un objet solide comme une union infinie de boules; d'autre part, certaines formes, et les modèles moléculaires de van der Waals en particulier, sont définies par une union de boules. Néanmoins, la question de savoir quel ensemble de boules utiliser pour approximer une forme est non trivial, de telle sorte que ce travail aborde deux problèmes liés. Pour les présenter, par conformation moléculaire, nous entendons un modèle défini par un ensemble fini de boules.

La premier problème, ou selection de diversité géométrique, consiste à choisir k conformations moléculaires parmi n, de façon à maximiser la diversité de l'ensemble choisi. Le second, ou approximation par défaut, consiste à approximer une molécule de n boules par $k < n$ boules.

Du point de vue théorique, nous montrons que les deux problèmes peuvent être traités avec une variante géométrique de max k-cover, les poids dépendant de la géométrie d'un arrangement surfacique ou volumique de sphères. La résolution de ces problèmes par un algorithme glouton permet d'avoir un facteur d'approximation borné inférieurement par $1 - 1/e$ dans certains cas. D'un point de vue appliqué, nous présentons une implémentation robuste de l'algorithme glouton pour l'approximation par défaut, laquelle incorpore (i) le calcul exact d'une triangulation de Delaunay dont les points ont des coordonnées qui sont des nombres algébriques de degré deux, (ii) le calcul exact de l'axe médian d'une union de boules, et (iii) une approximation certifiée du volume d'une union de boules. En particulier, nous montrons que des approximations précises de modèles moléculaires peuvent être obtenues en utilisant un nombre de boules 100 fois inférieur au nombre d'atomes, une propriété particulièrement séduisante pour la simulation d'environnement protéique dense.

Mots-clés : Optimisation géométrique, approximation géométrique, collections de boules, approximation par défaut, conformations moléculaires

Contents

1 Introduction

1.1 Modeling with Balls

Modeling complex 3D shapes is commonplace in science and engineering, and simple primitives such as balls play a central role in this process, for two reasons. On the one hand, the medial axis transform (MAT) allows representing a shape as a collection of balls [Ser82], usually infinite, so that sub-sampling such balls naturally yield approximations. On the other hand, (hierarchical) models represented by balls are ubiquitous, for example in molecular modeling, but also in robotics, computer graphics and CAGD, where bounding sphere hierarchies provide an elegant way to perform fast and numerically reliable collision detection. In this context, this paper addresses the following two problems, which we phrase using the molecular modeling terminology, even though the semantics of the balls might be different:

Conformational selection. Given a set of n conformations of a molecule represented by a collection of balls, select a *diverse* ensemble consisting of $s < n$ conformations.

Inner approximation. Given a (molecular) model consisting of n balls, provide an accurate volumetric approximation of this model using $s < n$ balls, contained in the original model.

These two problems are actually connected to a variety of research veins, namely (i) geometric approximation algorithms for 3D shapes, (ii) medial axis constructions and Voronoi diagrams, (iii) (geometric) approximation algorithms in general and max k-cover in particular, (iv) robust geometric software development, and (v) applications in structural biology. We now briefly comment on recent work in these directions.

As already mentioned, the problem of approximating 3D shapes is related to the medial axis transform (MAT). The particular case of a shape bounded by a smooth surface motivated the introduction of the MAT approximation using medial balls centered on specific Voronoi vertices called *poles* [AK00], an idea later re-used to approximate a shape bounded by a triangulated surface [AAK+09, SKS12]. This MAT approximation was also used for the *sphere-tree* construction [BO04], a representation to perform hierarchical object modeling and collision detection, and to improve the grasping quality in robotics [PAD10]. For a shape with smooth boundary, the previous MAT approximation typically comes with a guarantee, namely the convergence of the Hausdorff distance between the input boundary and that of the approximation. In a broader context, the problem of approximating a bounded open set has also been investigated recently. In [GMPW09] the authors introduce the scale-axis transform, which consists of scaling forth and back medial balls, so as to simplify a shape representation. It is worth noticing that all the works just mentioned rely on Voronoi diagrams, generally for the Euclidean distance, but also for a multiplicative distance in [GMPW09]. Consequently and from an implementation perspective, geometric algorithms from the Computational Geometry Algorithms Library [cga] (CGAL), but also number types from the `LEDA` [MN99] and `CORE` [KLPY99] libraries play a key role to implement such algorithms.

Our problems are also related to approximation algorithms in general, and greedy strategies in particular. As we shall see, of particular interest for both problems is max k-cover, which cannot be approximated within a ratio of $1 - 1/e + \varepsilon$ unless $\mathbf{P} = \mathbf{NP}$ [Fei98].

Last but not least, our incentive to tackle these two problems comes from computational structural biology, whose ultimate goal is to unravel the relationship between the structure and the function of macro-molecules and macro-molecular machines. Originating with the work of Richards [LR71], molecular models represented as collections of van der Waals (vdW) balls and associated affine Voronoi diagrams have been instrumental to describe atomic packing properties [MJLC87, MLJ+87], to compute and decorate molecular surfaces [Con83, AE96], to ex-

hibit correlations between structural and biological - biophysical properties of protein interfaces [BCRJ03, MDBC12], to select diverse conformational ensembles for mean field theory based docking algorithms [LSB+11], or to find entrance / exit passages to active sites [YFW+08]. More recently, compoundly weighted Voronoi diagram [CD10, DDC12] proved instrumental to assess the reconstruction of molecular machines involving up to circa 450 polypeptide chains [ADV+07], these reconstructions being plagued with uncertainties on the shapes and positions of the proteins. In fact, the inner covering problem is directly related to the design of simplified yet accurate protein models, in particular in the perspective of simulating crowded whole cellular environments [ME10, Goo09] — see also the beautiful illustrations of D. Goodsell[1].

1.2 Contributions and Paper Overview

Our two problems are concerned with shapes represented as a union of balls. These balls may have a molecular semantics, or may be carry a purely geometric meaning, e.g. in the MAT of any 3D shape. In using balls, a key advantage is that the structure of the medial axis of the union of these balls is known exactly [AM97, AK01]: it is actually coded by the α-shape of the input balls [Ede92], and the Voronoi diagram of the points found on the boundary of the union.

In this context, our contributions are threefold. On the theoretical side, we demonstrate that for both problems, a geometric generalization of max k-cover applies, with weights depending on the cells of a surface or volumetric arrangement. In particular, it is shown that the $1 - 1/e$ bound known in combinatorial optimization applies in some cases—but not all. From a geometric approximation perspective, our results depart from previous work since we focus on an approximation guarantees obtained with a finite set of balls rather than asymptotically. Second, on the software development side, we present an effective implementation for the inner approximation problem, based on state-of-the art geometric software. Finally, for the inner approximation problem, we present accurate coarse-grain protein models using a number of balls 100 times smaller than the number of atoms.

2 Geometric Optimization Problems for Balls

In the sequel, a *molecule* stands for a collection of balls. For the conformational selection, the molecule is termed a *conformer* when the relative position of the balls may change. The generic term *conformation* refers to a molecule or to a conformer, and we shall manipulate a collection of conformations $\mathcal{C} = \{C_i\}_{i=1,\ldots,n}$. The 3D *domain* spanned by the conformations is the union of their defining balls, and is denoted $\mathcal{F}_\mathcal{C} = \cup_{C_i \in \mathcal{C}} C_i$. For the inner approximation, we consider a single molecule. The set $\mathcal{C}$ refers to its constituting balls, and the domain $\mathcal{F}_\mathcal{C}$ is the union of these balls.

2.1 Geometric pre-requisites

Volume and surface decompositions. The spheres bounding the balls of a collection of conformations induce two decompositions: a decomposition of the volume occupied by the balls; and a decomposition of each sphere into spherical patches. More precisely, the decomposition of the volume $\mathcal{F}_\mathcal{C}$ induced by the spheres is called a *volumetric arrangement* (or *volumetric decomposition*). This arrangement consists of a collection of cells $\mathcal{A} = \{A_i\}$ such that the interior of each cell is connected. Each such cell is bounded by $2D$ cells, called surface patches, found on the spheres bounding the balls. On a given sphere, these patches are induced by the

[1] `http://mgl.scripps.edu/people/goodsell/books/MoL2-preview.html`

intersection circles with neighboring spheres. The collection $\mathcal{P} = \{P_i\}$ of all such patches defines a *surface arrangement* (or *surface decomposition*). A surface patch of a sphere which is not contained in any other ball is called *exposed*, and so is its supporting sphere. The collection of all such patches makes up the boundary $\partial\mathcal{F}_\mathcal{C}$ of the domain $\mathcal{F}_\mathcal{C}$. Note that each patch is bounded by circle arcs which are themselves delimited by points (generically) found at the intersection of three spheres. See Fig. 1 for a 2D illustration.

Medial axis of a collection of balls. The medial axis of the domain $\mathcal{F}_\mathcal{C}$ consists of the points having several neighbors on $\partial\mathcal{F}_\mathcal{C}$. Let a *boundary point* of $\mathcal{F}_\mathcal{C}$ be a 0-cell of $\partial\mathcal{F}_\mathcal{C}$. As proved in [AK01] and illustrated on Fig. 2, the MA consists of so-called singular simplices of the α-complex for $\alpha = 0$, together with a subset of the Voronoi diagram of the boundary points located within the α-shape.

2.2 Generic Geometric Optimization Problems

Conformational selection. We shall be concerned with two classes of combinatorial optimization problems. To state them from a combinatorial viewpoint (see section 3.1 and 3.2 for the connexion with conformations), assume we are given a base set $\mathcal{U} = \{U_i\}_{i=1,\ldots,m}$ of interior disjoint *cells* (think cells of the volume or surface arrangement), and a collection of *sets* $\mathcal{C} = \{C_i\}_{i=1,\ldots,n}$ called the *pool* (think conformations), where each set is a union of cells. For a subset $\mathcal{S} \subset \mathcal{C}$, denote $\mathcal{F}_\mathcal{S} = \cup_{C_j \in \mathcal{S}} C_j$ the union of the sets in $\mathcal{S}$. Cells and sets shall be subsets of $\mathbb{R}^3$, so that the inclusion of a cell U_i in a set C_j is naturally defined.

For the first class of problem, assume we are given a weight function w, i.e. a real valued function defined over the cells. Let $\binom{\mathcal{C}}{s}$ stand for the subsets of $\mathcal{C}$ of size s. We define:

Problem 1. *Given a weight function w, find a subset $\hat{\mathcal{S}}$ of $\mathcal{C}$ of size s, called the* selection, *such that:*

$$\hat{\mathcal{S}} = \arg\max_{\mathcal{S} \in \binom{\mathcal{C}}{s}} w(\mathcal{S}), \text{ with } w(\mathcal{S}) = \sum_{U_i \subset \cup_\mathcal{S} C_j} w(U_i). \tag{1}$$

For the second class of problems, assume the weight function depends not only on the cells of the decomposition, but also on the selection $\mathcal{S}$, which we denote $w_\mathcal{S}(U_i)$. We wish to solve:

Problem 2. *Given a weight function $w_\mathcal{S}$, find a subset $\hat{\mathcal{S}}$ of $\mathcal{C}$ of size s, called the* selection, *such that:*

$$\hat{\mathcal{S}} = \arg\max_{\mathcal{S} \in \binom{\mathcal{C}}{s}} w(\mathcal{S}), \text{ with } w(\mathcal{S}) = \sum_{U_i \subset \cup_\mathcal{S} C_j} w_\mathcal{S}(U_i). \tag{2}$$

Inner approximation. Consider a collection of n balls making up the domain $\mathcal{F}_\mathcal{C}$, that is $\mathcal{F}_\mathcal{C} = \cup_{B_i \in \mathcal{C}} B_i$, and let $\mathrm{Vol}(D)$ denote the 3D volume of a domain D. We introduce the following covering problem, phrased as an inner approximation problem:

Problem 3. *Find a collection $\mathcal{S}$ of s balls, such that $\mathcal{F}_\mathcal{S} \subset \mathcal{F}_\mathcal{C}$ and $Vol(\mathcal{F}_\mathcal{C} \backslash \mathcal{F}_\mathcal{S})$ is minimized.*

This problem is more constrained than the minimization of the volume of the symmetric difference $\mathrm{Vol}(\mathcal{F}_\mathcal{S} \Delta \mathcal{F}_\mathcal{C})$. Yet, the inner approximation problem is natural since it requires using balls centered on the medial axis of the input domain, since any other ball would be contained within a maximal ball.

Complexity issues. Our problems are intimately related to *max k-cover*. Given a set $\mathcal{U}$ of n points, and a collection $\mathcal{C}$ of subsets of $\mathcal{U}$, max k-cover aims at selecting k subsets from $\mathcal{C}$ so as to maximize the number of points from $\mathcal{U}$ which get covered [GJ79, Fei98]. (In the literature, this problem is sometimes called set cover [FG89]. To avoid confusion, we consider that the set cover problem aims at minimizing the number of sets in $\mathcal{C}$ to cover at least k elements from $\mathcal{U}$.) We note that the classical max k-cover is a special case of Problem 1 with function w assigning a unit weight to all cells. Since max k-cover is a **NP** complete problem, a polynomial time solution both $|\mathcal{C}|$ and s cannot be expected. However, the problem is in **P** for a fixed s since all subsets of size s can be probed. But this brute force method is doomed to fail even for moderate s, which calls for alternate strategies, the greedy strategy being the most natural one.

2.3 The Greedy Strategy

The greedy strategy consists of s iterations, the kth step consisting of selecting the C_j maximizing the weight of the union of the C_j. The selection achieved at step k shall be denoted $\mathcal{S}_k$. Unfortunately, the selection $\mathcal{S}_s$ may not realize the optimum solution, as illustrated by the simple example of Fig. 3. Thus, the performance assessment of greedy relies on the worst-case ratio between the solution returned and the optimal one. For max k-cover, this ratio is known to be of $1 - 1/e$, and is tight [CFN77, NWF78, FG89, Fei98].

3 Conformational Selection and the Greedy Strategy

We first examine problems concerned with the selection of a subset of conformers maximizing some *diversity* criterion.

3.1 Volumetric Decompositions and the Greedy Strategy

Problem 1 from volumetric decomposition. Consider the base set $\mathcal{A}$ whose cells are those of the $3D$ arrangement. In Eq. (1), let w be some general function defined on the cells of the volumetric decomposition. The weighting scheme is called non-negative provided all weights are ≥ 0—the most natural example being the standard Euclidean volume.

The approximation ratio of the greedy strategy and its optimality are usually proved in the uniform weight case [CFN77, NWF78, FG89, Fei98]. We shall prove the following

Theorem 3.1. *Consider a volumetric decomposition with non-negative weights. For Problem 1, the greedy approach has an approximation ratio of* $1 - (1 - 1/s)^s > 1 - 1/e$.

We shall use the following notation. The conformer selected at the k^{th} step is denoted C_k, and the weight of the optimum set of conformers OPT. Also, let $w^*(C_k)$ be the sum of the weights of the new elements in C_k that have not been covered in $C_j, 1 \leq j < k$ (i.e. the weight increment at step k). We start with a lemma needed to prove theorem 3.1. The proof of the lemma follows the usual one for max k-cover, but we include it for two reasons: first, it helps spotting the condition on the weight-functions w (the positivity is mandatory); second, in section 4.2, we shall re-use the skeleton of this lemma to prove a result of the greedy strategy for inner covering, with respect to the total volume of $\mathcal{F}_{\mathcal{C}}$ instead of the optimum.

Lemma 3.2. *For* $1 \leq k \leq s$, *the following holds:*

$$w^*(C_k) + \frac{1}{s} \sum_{j=1}^{k-1} w^*(C_j) \geq \frac{OPT}{s}. \tag{3}$$

Proof. At the k^{th} step, we select C_k that maximizes the weight of the new cells U_i being covered. Because the cells selected up to step $k-1$ may cover cells which are not covered by OPT, the weight of the cells that are covered by the optimum solution but not yet covered by the $(k-1)$ is at least

$$OPT - \sum_{j=1}^{k-1} w^*(C_j) \tag{4}$$

Since w is non-negative, the union-bound property states that for any collection of conformers $C_1, \ldots, C_p$, one has $w(C_1 \cup \cdots \cup C_p) \leq \sum_{i=1,\ldots,p} w(C_i)$. Since all the cells involved in Eq. (4) are covered by the optimum set of conformers, by the union-bound property, there must exist one conformer,not yet selected, that covers these new cells with total weight at least

$$\frac{1}{s}\left(OPT - \sum_{j=1}^{k-1} w^*(C_j)\right). \tag{5}$$

Since C_k maximizes the weight of the new cells being covered, we must have

$$w^*(C_k) \geq \frac{1}{s}\left(OPT - \sum_{j=1}^{k-1} w^*(C_j)\right). \tag{6}$$

Rearranging completes the claim.

□

Remark 1. *The non-negativity assumption is critical in the proof of Lemma 3.2. As a counter-example, consider the sets $C_1 = \{e1, e2\}, C_2 = \{e2, e3\}$ with $w(e1) = w(e3) = 1$ and $w(e2) = -1$. The union-bound fails for $w(C_1 \cup C_2)$. This remark is of particular interest in bio-physics, where atoms are decorated with physical, chemical or biological properties. For example, a weighting function that would take into account the electrostatic properties, which may be negative, would preclude the application of the previous lemma.*

Using Lemma 3.2, the proof of Thm. 3.1 goes as follows:

Proof. Multiplying the inequality obtained in the previous lemma by $(s-1)/s$ and adding to the inequality for step two, we get

$$w^*(C_1) + w^*(C_2) \geq \left(1 + \left(\frac{s-1}{s}\right)\right)\frac{OPT}{s}$$

We multiply this equation again by $\left(\frac{s-1}{s}\right)$ and add to the equation for step three, and so on. We get the following,

$$\sum_{j=1}^{k} w^*(C_j) \geq \left(1 - \left(\frac{s-1}{s}\right)^k\right) OPT$$

For $k = s$, we get,

$$\frac{\sum_{j=1}^{s} w^*(C_j)}{OPT} \geq \left(1 - \left(\frac{s-1}{s}\right)^s\right)$$

The left hand side is the ratio of the weight of the subset of $\mathcal{C}$ chosen by the greedy approach and the optimum solution i.e. that approximation factor and hence we have the above theorem. The fact that the above ratio is greater than $1 - \frac{1}{e}$ for all s is a trivial exercise. □

We now prove that the bound of the previous is tight (see also Fig.4):

Theorem 3.3. *The greedy approach cannot perform better than* $1-(1-1/s)^s$.

Proof. Fix a given s. We shall construct an example where the greedy approach can achieve an approximation ratio arbitrarily close to $1-(1-\frac{1}{s})^s$.

Let

$$\mathcal{A} = \{A_i\}_{i=1,\ldots,(s^2+s)}$$

$$\forall i,j \text{ s.t. } 0 \leq i < s, 1 \leq j \leq s,\ w(A_{i.s+j}) = \frac{1}{s^2}\left(\frac{s-1}{s}\right)^i$$

$$\forall j \text{ s.t. } 1 < j \leq s,\ w(A_{s^2+j}) = \frac{1}{s}\left(\frac{s-1}{s}\right)^s - \epsilon$$

The conformers are defined as follows

$$\mathcal{C} = \{C_i\}_{i=1,\ldots,2s}$$

$$\forall i \text{ s.t. } 1 \leq i \leq s,\ C_i = \bigcup_{j=(i-1).s+1}^{i.s} A_j$$

$$\forall i \text{ s.t. } \leq i \leq s,\ C_{s+i} = \bigcup_{j \equiv i \pmod{s}} A_j$$

Simple calculations lead us the following total weights for the conformers

$$\forall 1 \leq i \leq s,\ w(C_i) = \frac{1}{s}\left(\frac{s-1}{s}\right)^{i-1}$$

$$\forall 1 \leq i \leq s,\ w(C_{s+i}) = \frac{1}{s} - \epsilon$$

The optimum choice of $\mathcal{S}$ with $|\mathcal{S}| = s$ is clearly $\{C_i\}_{i=s+1,\ldots,2s}$ with total weight $1-s\epsilon$, whereas the greedy method would choose $\{C_i\}_{i=1,\ldots,s}$, with a maximum weight of $1-(1-\frac{1}{s})^s$, giving an approximation factor is arbitrarily close to $1-(1-\frac{1}{s})^s$. □

3.2 Surface Decompositions and the Greedy Strategy

Problem 2 from surface decomposition. Consider the base set $\mathcal{P} = \{P_i\}$ whose cells are those of the $2D$ arrangements induced on each sphere by the intersection circles with all the other spheres, computed e.g. with the algorithm from [CL09]. Special cells of this arrangement are those which are exposed, i.e. contribute to the boundary of the union of balls. Focusing on these patches yields an instantiation of Problem 2, the dependence upon the selection $\mathcal{S}$ consisting of discarding the patches which are not exposed with respect to the selection. For example, $w_{\mathcal{S}}(P_i)$ = surface area of patch P_i iff P_i is found on the boundary of the union $\mathcal{F}_{\mathcal{S}}$, and 0 otherwise.

Interestingly, maximizing the boundary surface of the selection is an indirect way to ascertain some diversity, since the overlap between conformers is minimized. Notice, though, that as opposed to the volume, the boundary surface area is not a monotonic function of the number of conformers. That is, for two selections S_1 and S_2 with $S_1 \subset S_2$, one has $volume(S_2) \geq volume(S_1)$, a property that may not hold for the boundary surface area.

Surface decomposition, boundary surface weight $w_{\mathcal{S}}$. For volumetric decompositions, the previous bound indicates that one is always above 63% $(1-1/e)$ from the optimum. Unfortunately, such a result does not hold for problem 2:

Observation 1. *Consider a surface decomposition. For Problem 2, the greedy approach may have a worst-case approximation ratio as bad as* $1/s^2$.

Proof. Consider a large ball B, and place s small non-intersecting balls $(B_1, \ldots, B_s)$ with their centers on the surface of B. The surface of each B_i is now divided into 2 patches. To the patch which lies inside B, we assign a weight of s. To each surface patch of B covered by some B_i, we assign a weight of $1+\epsilon$. All other surface patches are assigned a weight of 0.

The greedy strategy would first pick B because it has the largest exposed weight of $s(1+\epsilon)$. Now picking any $s-1$ of the B_i's would leave us with an exposed weight of only $s(1+\epsilon)-(s-1)(1+\epsilon) = 1+\epsilon$. On the opposite, a selection of the s smalls balls would have given us total exposed surface weight of s^2. This approximation factor is arbitrarily close to $1/s^2$. □

4 Inner Approximation and the Greedy Strategy

We establish two results for the inner approximation: first the performance of greedy with respect to the optimal solution, and second, with respect to the total volume $\mathrm{Vol}(\mathcal{F}_\mathcal{C})$.

4.1 Approximation Factor

As discussed when introducing problem 3, the inner approximation problem requires using balls centered on the medial axis. But the medial axis is a cell complex with two dimensional faces, so that one has an infinite collection of balls to choose from. To circumvent this difficulty, consider the following classical lemma [Ber87]:

Lemma 4.1. *Consider two intersecting spheres* Σ_1 *and* Σ_2 *in 3D, and define their convex linear combination, namely* $\Sigma_\lambda = \lambda\Sigma_1 + (1-\lambda)\Sigma_2$, *with* $\lambda \in [0,1]$. *The ball bounded by* Σ_λ *is contained in the union of the balls bounded by* Σ_1 *and* Σ_2.

Denote B_p^* a maximal ball centered on a vertex p of the medial axis, and let $\mathcal{V}$ be the set of vertices of the medial axis of $\mathcal{F}_\mathcal{C}$. By the structure theorem of the medial axis of a union of balls [AK01], this set is finite. We shall use this set to run the greedy algorithm, based on the following:

Observation 2. *The input domain* $\mathcal{F}_\mathcal{C}$ *satisfies*

$$\mathcal{F}_\mathcal{C} = \bigcup_i B_i = \bigcup_{v\in\mathcal{V}} B_v^*. \tag{7}$$

Proof. We shall prove that any maximal ball B_p^* is contained in the union of at most three balls centered on vertices from $\mathcal{V}$. Omitting the trivial case of a singular vertex of the medial axis, we first note that there are three cases to be analyzed, namely when p belongs to a singular edge of the medial axis, when it belongs to a singular triangle, or when it belongs to a (possibly clipped) Voronoi face f.

Case 1. This is exactly the case covered by lemma 4.1. In this case, the portion of the pencil contains the intersection circle between the two spheres defining the singular edge.

Case 2. The second case contains two sub-cases, namely when p lies in the interior of a Voronoi edge, and when p lies in the interior of the Voronoi facet f. The first sub-case is again the case of lemma 4.1 — all the spheres in the portion of the pencil contain the three boundary points defining the Delaunay triangle dual of the Voronoi edge in question. For the second one: let c be

any Voronoi vertex of f belonging to $\mathcal{V}$, let L be the ray emanating from c and passing through p, and let d be the intersection point between L and the boundary ∂f of f. Point d belongs to either a Voronoi edge or to an α-shape edge (if the Voronoi facet is a clipped Voronoi facet in the medial axis). Call e and f the endpoints of this 1-cell of the medial axis. Now, by lemma 4.1, one has $B_d^* \subset B_e^* \cup B_f^*$ and similarly $B_p^* \subset B_c^* \cup B_d^*$. Thus, $B_p^* \subset B_c^* \cup B_e^* \cup B_f^*$.
Case 3. Amenable to the analysis carried out for Case 2.

Thus, since any maximal ball is contained in the union of at most three balls centered at vertices from $\mathcal{V}$, the claim holds. □

The corollary of the previous observation is that the balls centered on the medial axis constitute the pool of candidate balls to choose from by the greedy strategy. Thus, theorem 3.1 applies, that is:

Theorem 4.2. *For Problem 3, the greedy approach based on the maximal balls centered on the vertices of the medial axis has an approximation ratio of* $1-(1-1/s)^s > 1-1/e$.

4.2 Worst Case Bound with respect to the Total Volume

Approximation bound. We generalize the previous result with respect to the total volume of the input domain:

Lemma 4.3. *Let* $V = Vol(\mathcal{F}_\mathcal{C})$ *be the volume of union of given* n *balls and let GREEDY be the volume of union of* s *balls selected by the greedy algorithm. These volumes satisfy:*

$$\frac{GREEDY}{V} \geq 1 - \left(1 - \frac{1}{n}\right)^s \tag{8}$$

Proof. In the proof of the approximation factor of greedy algorithm for volumetric decomposition given in lemma 3.3, note that it is valid for any solution and not only the optimum solution, i.e. no property of the optimum solution is required. Thus we replace the optimum solution by a solution selecting the given n balls. Thus we get the following equation.

$$w^*(C_k) \geq \frac{1}{n}\left(V - \sum_{i=1}^{k-1} w^*(C_i)\right)$$

where C_i is the i^{th} ball selected by the greedy algorithm, and $w^*(C_k)$ is the new volume of C_k not covered by any of C_i, $1 \leq i < k$. Solving it in the manner similar to that used in the proof of Them 3.3 yields:

$$GREEDY = \sum_{i=1}^{s} w^*(C_i) \geq V \cdot \left(1 - \left(1 - \frac{1}{n}\right)^s\right)$$

□

Tight example. Consider n balls of same radii. Then greedy algorithm would select any s balls out of it. This would contribute a volume equal to s/n times the total volume.
Also note that

$$\frac{s}{n} = 1 - \left(1 - \frac{s}{n}\right) \approx 1 - \left(1 - \frac{1}{n}\right)^s$$

for very large values of n. In fact, this is the best that can be done, i.e. no algorithm can approximate union of n balls with approximation factor greater than s/n in worst case.

5 Inner Approximation: Software and Experiments

We focus on the inner approximation since the implementation of the conformational selection merely requires a robust algorithm to compute the volume of a union of balls in the volume case, and an algorithm to compute arrangements of all types of circles on a sphere in the surface case.

5.1 Greedy Algorithm

Basic algorithm. The input consists of a collection of balls, and of a selection size s or a target ratio τ (the volume of the selection divided by the volume of the input balls should be at least τ). The output consists of an ordering of the selected balls, together with the increment in volume associated to each ball. We also report the Betti numbers along the selection, computed with the algorithm from [DE95].

The algorithm consists of iteratively selecting the ball providing the best volume increment, selected from a priority queue containing all candidates. Upon selecting ball say B_i, we recompute the volume increments of all candidate balls intersecting B_i.

Imposing connectedness of the selection. For selected applications, the domain $\mathcal{F}_{\mathcal{S}}$ should be connected: for example, the selection associated to a connected molecule should also be connected. To meet this constraint, the following heuristic is used. Let $\mathcal{S}_k$ be the selection upon termination, and consider the exposed balls i.e. the balls contributing to the boundary $\partial\mathcal{F}_{\mathcal{S}}$. Split these balls into two groups L and L^c, namely the largest component (in number of exposed balls), and the remaining ones. We aim at connecting L to one of the connected components of L^c. To do so, using the Delaunay triangulation of the centers of the balls in $L \cup L^c \cup (\mathcal{S} \backslash \mathcal{S}_k)$, we compute the shortest path joining a center of a ball in L to a center of a ball in L^c. This shortest path uses centers of balls in $\mathcal{S} \backslash \mathcal{S}_k$, which are added to the section. This process is iterated until one connected component remains.

5.2 Geometric Objects

The previous algorithm involves elaborate geometric objects, which we present now by following the flow of the algorithm, mentioning the CGAL[2] classes used and their template parameters when appropriate.

The Delaunay triangulation DTB of the input balls, and the associated α-shape. Following classical usage, we call K the kernel used to instantiate the CGAL classes `Regular_triangulation_3` and `Alpha_shape_3`. Two options for K are discussed below.

The Delaunay triangulation DTV of the boundary points of $\partial\mathcal{F}_{\mathcal{S}}$. Two difficulties are faced to construct DTV. First, more than three co-planar points are generic in DTV [AK01]. Second, since a boundary point is found at the intersection of three input spheres, its coordinates are degree two algebraic numbers. We therefore store these points using the CGAL spherical kernel `Spherical_kernel_3` [CCLT09], instantiated with K. The two options for K, referred to as the *inexact* and the *exact* kernels in the sequel, are:

- `Exact_predicates_inexact_constructions_kernel`, the underlying number type (NT) to store the coordinates of the boundary points being a `double`.

[2] `http://www.cgal.org`

- `Exact_predicates_exact_constructions_kernel_with_sqrt`, the underlying number type to store the coordinates being either `CORE::Expr` or `LEDA::real`. Additionally, a map is used to associate a singular or regular facet from the α-shape of DTB to each boundary point.

To handle these difficulties, we implemented a dedicated kernel denoted `DTV_kernel`, defining a new point type for the boundary points. This kernel is actually templated by two parameters:

- First, a ball identifier, used to record the three input spheres defining a boundary point. These identifiers are used to handle the aforementioned special cases, so as to avoid the numerical calculation of a predicate whose sign can be inferred from the fact that the input points lie on a set of known input spheres. Practically and since an input ball corresponds to a vertex of the α-shape of DTB, the vertex handle of the α-shape is taken as identifier.

- Second, a number type used to represent the coordinates of the boundary points, the two options being the NT associated to the aforementioned inexact and exact kernels.

One comment is in order about the Voronoi diagram DTV^* dual of DTV, since medial ball associated to selected Voronoi vertices are used by greedy. With the inexact kernel, the input points of DTV are approximations of the exact boundary points, since the degree two algebraic number get converted to doubles. For these points, the combinatorial structures of DTV and DTV^* are exact (exact predicates are used), but the embedding of the Voronoi vertices of DTV^* is inexact (inexact constructions are used). With the exact kernel, the input points of DTV are exactly the boundary points. Moreover, the embedding of the Voronoi vertices is exact (exact constructions are used).

The medial-axis of the union of input balls. We store the medial axis as a container of polygons, possibly degenerate for singular vertices and edges of the α-shape [AK01]. Our polygon class inherits inherit from the CGAL class `Polygon_2` (embedded in 3D), instantiated with the kernel K. It offers new features, in particular the computation of the maximal ball centered at a point of the polygon. Such a ball has a center which is a `Point_3` from K, and a squared radius whose type is `NT`.

The candidate balls. Following the results of section 4.1, the candidate balls used are only centered on the vertices of the medial axis. Such balls are associated with the medial axis, as just discussed.

The volume of the selected balls. Computing the volume of a union of balls is a difficult problem, from a combinatorial, but also numerical standpoint—inverse trigonometric functions are involved. We use our certified algorithm [CKL11] which returns an interval certified to contain the exact volume. More precisely, due to the impossibility to obtain a volume as an exact number type, whatever the kernel used (exact, inexact), the centers and radii of the candidate balls are converted to doubles. These balls are input to our algorithm, which requires two template parameters: the number type of the output (double or interval), and the level of exactness used to compute the constructions involved in the volume computation, namely the coordinates of Voronoi vertices, and boundary points of the union of the selected balls. Following the discussion in [CKL11], the three options are referred to as (faster, ck_pt_exact and all_exact). Practically, we use the pair (double, faster) for the inexact kernel, and (interval, all_exact) for the exact kernel.

5.3 Results

Dataset. As test set, we used the 96 protein - protein complexes from [LCJ99]. The complexes are of high biological interest (all of them are coupled to well identified biological processes). The number of atoms lies in the range $[1008, 13214]$, with a median of 3757.

Performances and robustness issues. The properties of predicates and constructions of the exact and inexact kernels has been discussed in section section 5.2. We compared the volume ratios obtained with these two options on a set of 10 protein complexes, and did not observe any difference before the third digit. For two examples discussed in detail below, these ratios are ~ 0.69 for the protein complex 3sgb with $r = 2.8$ and 20 selected balls, and ~ 0.64 for the immunoglobulin 1igt with $r = 2.8$ and 103 selected balls. For running times, we compared the execution time for the construction of DTB, for DTV and for the medial axis. The selection itself was excluded since from the timing, as also noticed in section 5.2, our volume computation algorithm uses `double` as number type. On the aforementioned 10 models, we observed that the exact kernel was on average about 150 times slower than the inexact one. For these two reasons — absence of obvious degeneracies and much better running time, the results reported in the sequel were computed with the inexact kernel.

Using the inexact kernel, it is observed that the running times for computing DTB and DTV are a mere order of magnitude slower than the CGAL ones [3] for the regular triangulation case (Fig. 5). These running times are naturally consistent with the fact that the geometric objects manipulated behave nicely for our molecular models: both the number of boundary points (Fig. 6) and the primitives of the medial axis (Fig. 7) are linear in the number of input balls.

Inner approximation guarantees. Intuitively, the ability of greedy to provide a good approximation relies on the possibility to choose large balls, which depends itself on two parameters. First, the topological complexity: the closest to a topological ball the domain $\mathcal{F}_{\mathcal{C}}$, the better. Second, the geometric complexity: the more convex the domain $\mathcal{F}_{\mathcal{C}}$, the better. Before commenting these properties on molecular systems, recall that in a vdW model, the radii of the balls vary in the range 1-2Å, and that only balls of atoms linked by covalent bonds intersect. Thus, for a vdW model, one expects the volume covered to vary linearly as a function of the selection size, which is exactly observed (Fig. 8, red curves). Now, enlarging the input balls by a quantity r_w, e.g. $r_w = 1.4$ to define a solvent accessible model, results in simplifying the topology of $\mathcal{F}_{\mathcal{C}}$. Thus, the larger r_w, the larger the candidate balls, and the better the ratio curve (Fig. 8 again). For a fixed budget of balls and a given expansion radius, e.g. $r_w = 1.4$, one also expect complex topologies, characterized by a high Euler characteristic, to yield more difficult problems. This trend is also observed (Fig. 9).

As for the incidence of the overall shape, a roughly convex system (Fig. 10(A) versus Fig. 11(A)) clearly yields more favorable volume ratio curves (Fig. 8(Top, Bottom)).

Coarse graining molecular models. The immunoglobulin (Ig) structure just used provides a good example, because of its non convexity, to test the algorithm to go beyond the inner covering problem. This model also exhibits a topological rather than geometric difficulty, since at the tip of the two arms each so-called *variable domain* has the topology of a filled torus (Fig. 11(A)). To this end, in a manner identical to the scale axis transform [GMPW09], we may enlarge the model by r_w, approximate it, and compare the balls obtained with the initial vdW

[3] `http://www.cgal.org/Manual/latest/doc_html/cgal_manual/Triangulation_3/Chapter_main.html#Subsection_39.6.1`

model. When implemented with $r_w = 2.8$, this heuristic yields a coarse-grain model exhibiting a discrepancy of about 3 atoms sticking out from the coarse grain model in the worst-case (Fig. 11(C)). Moreover, the topology of the variable domains has also been preserved. Thus, one gets an accurate model for a number of balls which has been divided by about 100 in this case. (In biophysics, a resolution of 5Å is the maximum which may be tolerated to say that a model has atomic resolution.)

6 Conclusion and Outlook

This paper studies two basic problems dealing with collection of balls, namely selecting a diverse set of molecular conformations, and providing an accurate inner approximation of a molecular model. Both problems are shown to be geometric versions of max k-cover, the weight function being a function of the geometry of the cells of a surface or volumetric arrangement, rather than being uniform as in the combinatorial setting. Yet, for the volumetric case, the approximation bound known in the combinatorial setting is preserved, provided that the weights are non-negative. The implementation of our algorithms hinges upon state-of-the-art software coupled to the CGAL library. In particular, this implementation involves the exact calculation of a Delaunay triangulation for points whose coordinates are degree two algebraic numbers, the intersection of the dual of this triangulation with the α-complex of the input balls, and the certified calculation of the volume of a union of medial balls. This implementation handles molecular models containing up to $O(10^5)$ atoms within minutes. For these reasons, we believe that our algorithms, the inner approximation in particular, should prove useful for a broad class of geometric approximation problems dealing with balls, in particular in the context of approximate medial axis transforms, where the focus has been so far on asymptotic properties—upon increasing the number of balls.

Yet, our work calls for further developments, both in the theoretical and applied directions. On the theoretical side, two challenging questions are of high interest. First, our greedy algorithms come with guarantees for the inner approximation problem. But monitoring the symmetric difference of the input domain and of the selection, or the Hausdorff distance between their boundaries is also clearly of interest. Second, constraining the geometric selection by topological criteria, e.g. prescribed Betti numbers, would also be of the highest interest. However, approximation problems aiming at accommodating both geometric and topological criteria are likely to be challenging—it has been shown that the so-called homology localization problem is **NP**-hard. In an applied vein and as mentioned in introduction, we believe that a key application of our algorithms will be the design of coarse-grain macro-molecular models, to investigate macro-molecular machines and simulate crowded environments within whole cells. But prior to undertaking these challenges, one will have to decorate our purely geometric coarse-grain models with bio-physical properties.

Acknowledgments. Michael Hemmer is acknowledged for his help with the CGAL kernels.

This work has partially been supported by the Computational Geometric Learning STREP project of the EC 7th Framework Programme (EC contract No. 255827).

References

[AAK+09] O. Aichholzer, F. Aurenhammer, B. Kornberger, S. Plantinga, G. Rote, A. Sturm, and G. Vegter. Recovering structure from r-sampled objects. In *Computer Graphics Forum*, volume 28, pages 1349–1360. John Wiley & Sons, 2009.

[ADV+07] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[AE96] N. Akkiraju and H. Edelsbrunner. Triangulating the surface of a molecule. *Discrete Applied Mathematics*, 71(1-3):5–22, 1996.

[AK00] N. Amenta and R.K. Kolluri. Accurate and efficient unions of balls. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 119–128. ACM, 2000.

[AK01] N. Amenta and R. K. Kolluri. The medial axis of a union of balls. *Comput. Geom. Theory Appl.*, 20:25–37, 2001.

[AM97] D. Attali and A. Montanvert. Computing and simplifying 2d and 3d continuous skeletons. *Computer Vision and Image Understanding*, 67(3):261–273, 1997.

[BCRJ03] R.P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins: Structure, Function, and Bioinformatics*, 53(3):708–719, 2003.

[Ber87] M. Berger. *Geometry I*, volume 1. Springer, 1987.

[BO04] G. Bradshaw and C. O'Sullivan. Adaptive medial-axis approximation for sphere-tree construction. *ACM Transactions on Graphics (TOG)*, 23(1):1–26, 2004.

[CCLT09] P. M. M. De Castro, F. Cazals, S. Loriot, and M. Teillaud. Design of the cgal spherical kernel and application to arrangements of circles on a sphere. *Computational Geometry: Theory and Applications*, 42(6-7):536–550, 2009.

[CD10] F. Cazals and T. Dreyfus. Multi-scale geometric modeling of ambiguous shapes with toleranced balls and compoundly weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, pages 1713–1722, Lyon, 2010. Also as INRIA Tech report 7306.

[CFN77] G. Cornuejols, M.L. Fisher, and G. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[cga] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[CKL11] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 38(1):1–20, 2011.

[CL09] F. Cazals and S. Loriot. Computing the exact arrangement of circles on a sphere, with applications in structural biology. *Computational Geometry: Theory and Applications*, 42(6-7):551–565, 2009.

[Con83] M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Crystallogr.*, 16(5):548–558, 1983.

[DDC12] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macro-molecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 80(9):2125–2136, 2012.

[DE95] C.J.A. Delfinado and H. Edelsbrunner. An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere. *Computer Aided Geometric Design*, 12(7):771–784, 1995.

[Ede92] H. Edelsbrunner. Weighted alpha shapes. Technical Report UIUCDCS-R-92-1760, Dept. Comput. Sci., Univ. Illinois, Urbana, IL, 1992.

[Fei98] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

[FG89] PC Fishburn and WV Gehrlein. Pick-and choose heuristics for partial set covering. *Discrete Applied Mathematics*, 22(2):119–132, 1989.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, NY, 1979.

[GMPW09] J. Giesen, B. Miklos, M. Pauly, and C. Wormser. The scale axis transform. In *ACM Symp. on Computational Geometry*, pages 106–115, 2009.

[Goo09] D. Goodsell. *The machinery of life*. Springer, 2009.

[KLPY99] V. Karamcheti, C. Li, I. Pechtchanski, and C. Yap. A core library for robust numeric and geometric computation. In *15th ACM Symp. on Computational Geometry, 1999*, pages 351–359, 1999.

[LCJ99] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285:2177–2198, 1999.

[LR71] B. Lee and F. M. Richards. The interpretation of protein structure: Estimation of static accessibility. *J. Molecular Biology*, 55:379–400, 1971.

[LSB+11] S. Loriot, S. Sachdeva, K. Bastard, C. Prevost, and F. Cazals. On the characterization and selection of diverse conformational ensembles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):487–498, 2011.

[MDBC12] N. Malod-Dognin, A. Bansal, and F. Cazals. Characterizing the morphology of protein binding patches. *Proteins: structure, function, and bioinformatics*, 80(12):2652–2665, 2012.

[ME10] S.R. McGuffee and H. Elcock. Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.*, 6(3):1–18, 2010.

[MJLC87] S. Miller, J. Janin, A.M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–656, 1987.

[MLJ+87] S. Miller, A.M. Lesk, J. Janin, C. Chothia, et al. The accessible surface area and stability of oligomeric proteins. *Nature*, 328(6133):834–836, 1987.

[MN99] K. Mehlhorn and S. Näher. *LEDA: a platform for combinatorial and geometric computing*. Cambridge University Press, 1999.

[NWF78] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[PAD10] M. Przybylski, T. Asfour, and R. Dillmann. Unions of balls for shape approximation in robot grasping. In *Proc. IEEE Conf. Intelligent Robots and Systems (IROS)*, pages 1592–1599, 2010.

[Ser82] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, London, UK, 1982.

[SKS12] S. Stolpner, P. Kry, and K. Siddiqi. Medial spheres for shape approximation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1234–1240, 2012.

[YFW+08] E. Yaffe, D. Fishelovitch, HJ. Wolfson, D. Halperin, and R. Nussinov. Molaxis: Efficient and accurate identification of channels in macromolecules. *Proteins*, 73(1):72–86, 2008.

7 Artwork

7.1 Theory

Figure 1 2D conformations, each consisting of four balls—first and fourth balls are common, and the induced surface and volume arrangements. The (two dimensional) volume occupied by the two conformations is decomposed into 19 cells (boxed numerals). The circled numerals feature the surface arrangement of the ball centered at a_1, based on intersections with neighboring balls.

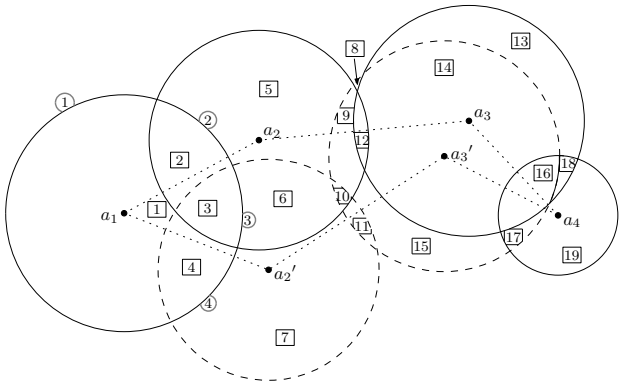
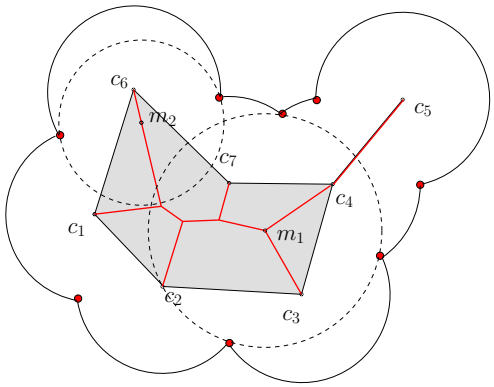


Figure 2 The medial axis transform for a union of balls, in 2D. The boundary points of the union of the seven balls are represented by red dots, while the medial axis is presented by red line-segments. Two maximal balls centered on the medial axis are presented in dashed circles (their centers are m_1 and m_2. Each such ball touches the boundary $\partial\mathcal{F}_\mathcal{C}$ in at least two points.



.

Figure 3 Greedy selection may not yield the optimal solution. Greedily selecting two sets out of $C_1, \ldots, C_5$ yields a score of 12 (selecting C_3 and then C_2), while the optimum is 14 (selecting C_4 and C_5). The shaded cells have the weights as indicated and the unshaded cells have null weights.

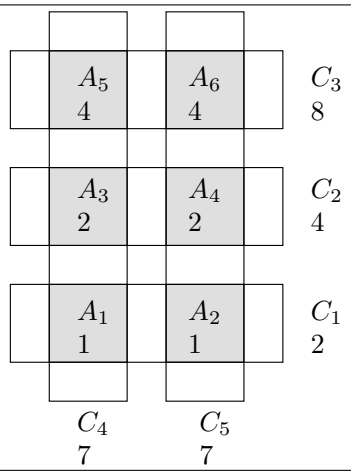
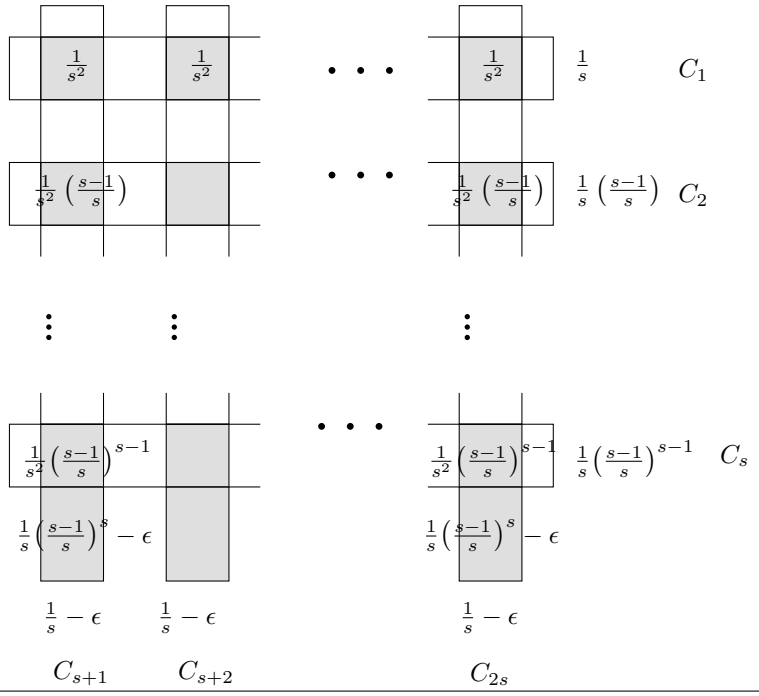


Figure 4 A tight example for the greedy strategy.



7.2 Running Times

Figure 5 Running times for the key steps of the inner approximation algorithm, as a function of the number of input balls. The models used are from [LCJ99]. (i) DTB of input balls (ii)DTV of boundary vertices (iii) Medial axis of the union of balls

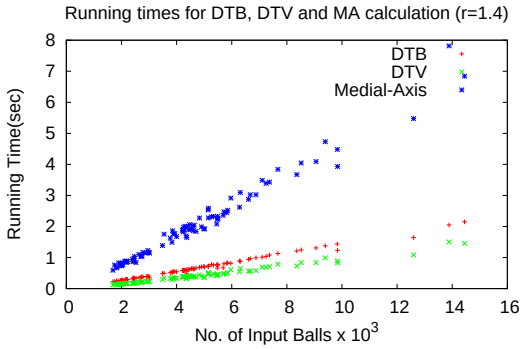


Figure 6 Number of boundary vertices as a function of the number of input balls.

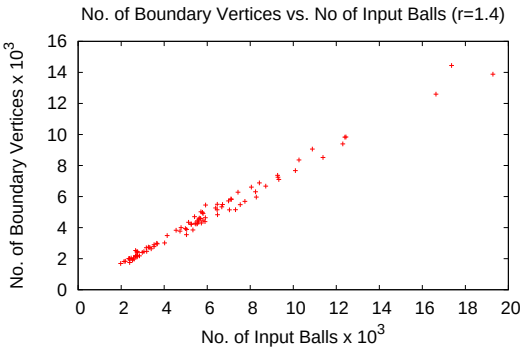
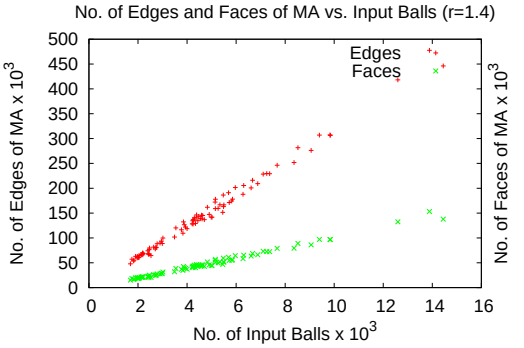


Figure 7 Number of faces of the medial axis as a function of the number of input balls.



7.3 Approximation Guarantees

Figure 8 Variation of volume ratios of each size of selection, wrt added radius of input balls: **Top** Protein complex of 1690 balls (PDB code 3sgb, see Fig. 10) **Bottom** Immunoglobulin of 10416 balls (PDB code 3sgb, see Fig. 11)

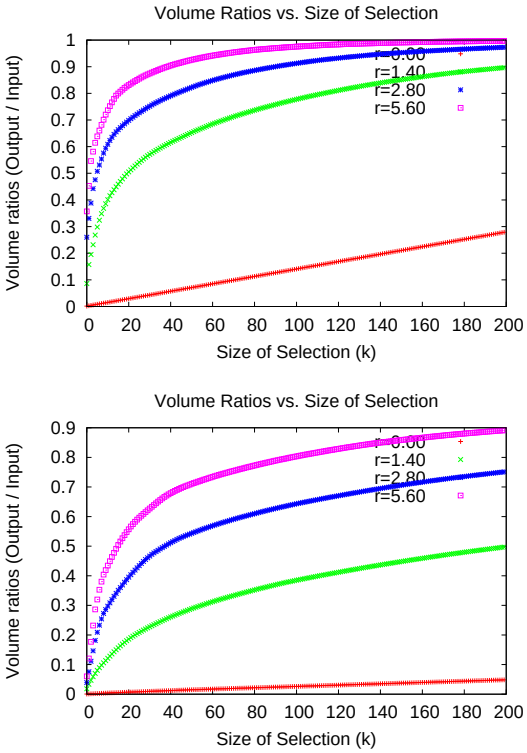


Figure 9 Volume ratio as a function of the Euler characteristic of the input model, for $r_w = 1.4$, and a budget of $s = 200$ balls. Each point represents a different input molecule.

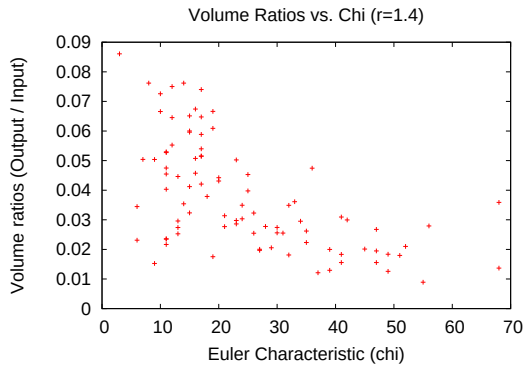


Figure 10 Coarse graining a molecular model: the example of a small globular protein complex (PDB id: 3sgb) (A) The atomic van der Walls models contains 1,872 atoms. **(B)** A coarse grain model of 20 balls, defined as the inner approximation of the atoms whose van der Waals radii have been expanded by $r_w = 2.8$. **(C)** The superimposition of both models.

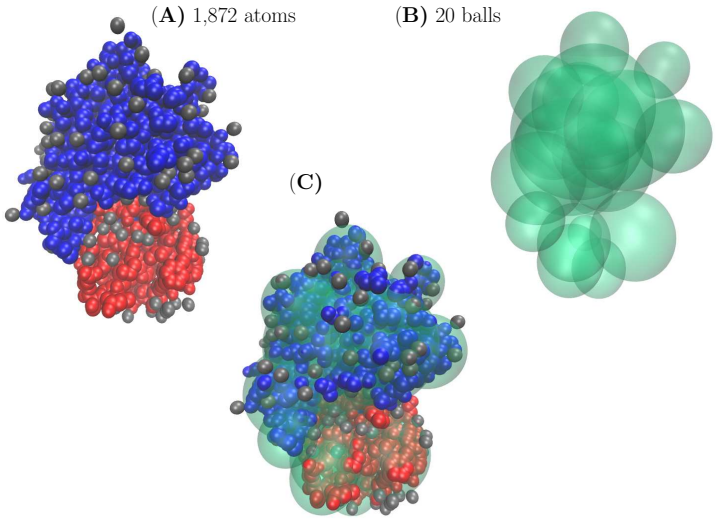
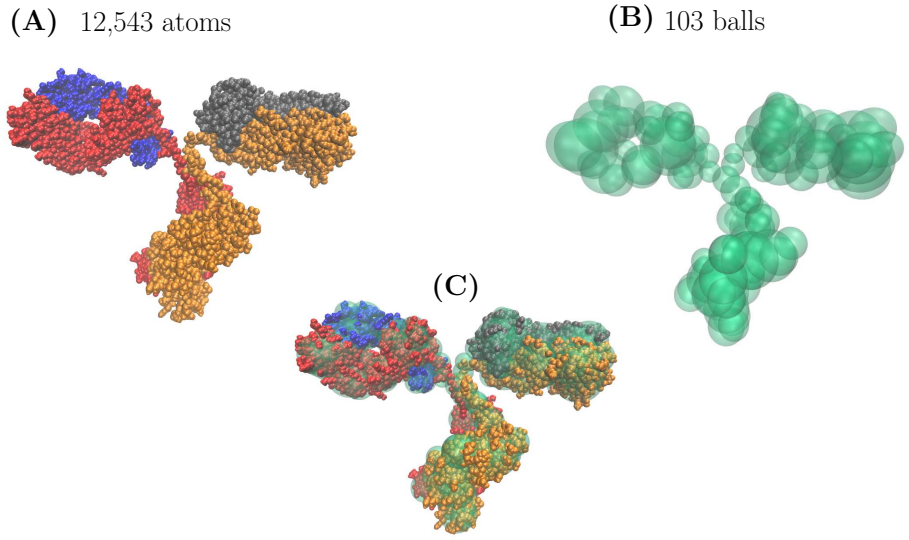
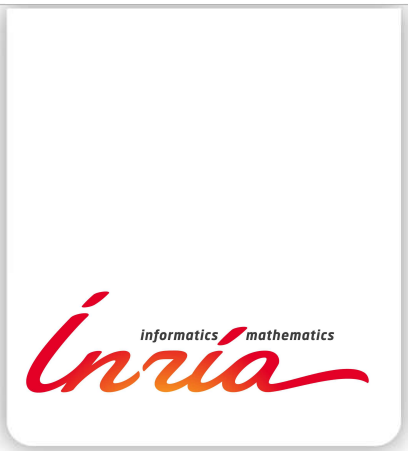


Figure 11 Coarse graining a molecular model: the example of an immunoglobulin (PDB id: 1igt). (A) The atomic van der Walls models. **(B)** A coarse grain model of 103 balls, defined as the inner approximation of the atoms whose van der Waals radii have been expanded by $r_w = 2.8$. Three balls were added to the greedy selection with $s = 100$ as explained in section 5.1 to force the connectivity. **(C)** The superimposition of both models. Note in particular that the hole of the *arm* on the left hand side is respected.





RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399